Mathématiques et sciences humaines
Mathematics and social sciences

154 | Été 2001
Analyse statistique implicative

Un algorithme de regroupements de modalités de variables en analyse implicative des données

Grouping together variables values: an algorithm in implicative analysis

Dominique Lahanier-Reuter

Édition électronique
URL : http://journals.openedition.org/msh/2847
DOI : 10.4000/msh.2847
ISSN : 1950-6821

Éditeur
Centre d'analyse et de mathématique sociales de l'EHESS

Édition imprimée
Date de publication : 1 mars 2001
ISSN : 0987-6936

Référence électronique
Dominique Lahanier-Reuter, « Un algorithme de regroupements de modalités de variables en analyse implicative des données », *Mathématiques et sciences humaines* [En ligne], 154 | Été 2001, mis en ligne le 10 février 2006, consulté le 20 avril 2019. URL : http://journals.openedition.org/msh/2847 ; DOI : 10.4000/msh.2847

Math. & Sci. hum., (39e année, n° 154-155, *2001*, p. 47-59)

UN ALGORITHME DE REGROUPEMENTS DE MODALITÉS DE VARIABLES EN ANALYSE IMPLICATIVE DE DONNÉES

Dominique LAHANIER-REUTER [1]

RÉSUMÉ — *Dans le cadre de l'analyse implicative de données, nous développons un algorithme original de regroupements de modalités de variables. Il s'agit de déchiffrer des lignes de force implicatives lorsque les modalités observées des variables étudiées sont en nombre élevé, par exemple lorsque l'une des variables est numérique.*

MOTS-CLÉS — Analyse implicative, Algorithme, Comparaison de distributions.

SUMMARY — Grouping together variables values: an algorithm in implicative analysis
We develop an algorithm aimed at gathering variable values, in the frame of implicative data analysis. This procedure is presented in the case of one quantitative variable.

KEYWORDS — Implicative analysis, Algorithm, Distribution comparison.

1. INTRODUCTION

Nous présentons un algorithme de regroupements de modalités observées de deux variables. Cet algorithme a pour but de mettre en évidence des lignes de force de type implicatif, au sens statistique du terme, dans le cas où les valeurs relevées des deux variables engagées sont soit numériques, soit en nombre élevé. Dans ce cas, la liste des implications statistiques entre différentes modalités des deux variables est souvent difficile à interpréter.

Nous commençons par présenter une recherche où le problème de la gestion des implications statistiques entre données est apparu. Nous développerons une solution à ce problème en exposant la procédure algorithmique élaborée tout en la faisant fonctionner sur l'exemple choisi.

[1] Équipe THEODILE, EA 1764, Université Charles-de-Gaulle Lille III, Domaine universitaire du «Pont de Bois», BP 149, 59653 Villeneuve d'Ascq cedex, e-mail : Dominiquereuter@aol.com

2. DESCRIPTION DE LA SITUATION EXPÉRIMENTALE ET INTÉRET DU RECOURS À L'ANALYSE IMPLICATIVE

La recherche contextuelle qui sert de cadre initial au développement vient d'un travail de thèse sous la direction de Régis Gras.

Travaillant sur les conceptions locales et pragmatiques du hasard, nous avons proposé trois conceptions différentes du hasard :

1. le hasard de l'aléatoire ;
2. le hasard de l'événement exceptionnel et improbable ;
3. le hasard des deux possibles [4].

Pour savoir si ces différentes conceptions étaient également mobilisées par des sujets d'âge et de parcours scolaires divers, nous avons recueilli des productions écrites de cent deux sujets, d'âge et de niveau de scolarité variés. On comptait dix-huit élèves d'une classe de CM1/CM2, âgés de 9 à 10 ans, puis vingt-quatre élèves d'une classe de terminale ES[2] , âgés de 17 à 20 ans, et enfin soixante étudiants de licence de Sciences de l'éducation, âgés de 20 à 45 ans. Chacun des sujets considérés est donc, soit CM pour les élèves de CM1/CM2, soit TES pour les élèves de la classe de terminale, soit Étudiant pour les étudiants de licence.

Il s'agissait de composer huit petits textes, à partir d'illustrations, avec pour seule contrainte d'utiliser le mot *hasard.*

Chacun des huit textes était classé – quand cela s'avérait possible – selon la conception du hasard que nous pensions y déchiffrer. Nous avons ainsi pu leur faire correspondre un triplet, constitué des nombres de textes s'inscrivant dans chacune des trois conceptions.

Les sujets mobilisaient-ils uniquement l'une de ces conceptions ou avaient-ils plutôt tendance à diversifier leurs positions ? Il était possible en effet d'obtenir aussi bien des triplets du type (8, 0, 0) que du type (3, 2, 3). Nous qualifions le premier triplet de « rigide », c'est-à-dire correspondant à un sujet qui avait, au long des huit textes, mobilisé une seule des trois conceptions, dans ce cas celle du hasard aléatoire. Le second triplet (3, 2, 3) caractérisait, pour nous, une certaine « souplesse » du sujet, puisqu'il rendait compte de la diversité des positions adoptées par le sujet au long de l'expérience.

Pour rendre compte de la « rigidité » ou de la « souplesse », nous avons élaboré une variable numérique qui permet de hiérarchiser les triplets : c'est l'entropie d'un triplet de somme constante égale à 8 (voir tableau 1). Une entropie faible traduit une rigidité des choix, une entropie élevée une souplesse.

---

[2] La classe de terminale ES est une option « économique et sociale ». Le programme de mathématiques de cette section est le plus riche en statistiques, comparativement à celui des autres sections.

Tableau 1. Entropie des différents triplets de somme 8

Triplets[3]	8,0,0	7,1,0	6,2,0	5,3,0	4,4,0	6,1,1	5,2,1	4,3,1	4,2,2	3,3,2
Entropie	0.000	0.543	0.813	0.954	1.000	1.061	1.298	1.405	1.500	1.561

En interrogeant le lien éventuel entre la qualité de la mobilité des choix (souplesse vs rigidité) et leur niveau scolaire, nous avons défini des relations entre modalités observées d'une variable numérique – l'entropie – et celle d'une variable qualitative – le niveau scolaire – tente d'en mesurer l'intensité. Le type d'analyse qui nous a alors semblé le plus pertinent est celui de l'analyse implicative de données et nous allons développer les raisons de ce choix.

3. L'ANALYSE IMPLICATIVE DE DONNÉES

Le tableau 2 de répartitions des 102 sujets selon l'entropie de leurs choix et leurs situations scolaires dans les catégories CM et TES, d'entropie 1,000 et 1,061, indique un déséquilibre presque symétrique des répartitions de l'entropie des élèves de CM et

Tableau 2. Distributions de l'entropie selon les catégories de sujets

	Niveau scolaire			
Entropie	CM	TES	Étudiants	Totaux
0.000	4	2	1	7
0.543	2	1	5	8
0.813	3	2	5	10
0,.954	6	6	4	16
1.000	0	0	8	8
1.061	0	0	6	6
1.298	2	6	8	16
1.405	1	1	11	13
1.500	0	3	5	8
1.561	0	3	7	10
Totaux	18	24	60	102

[3] Les triplets sont à lire aux permutations près.

des étudiants. Ceci suggère une répartition des élèves de CM en un groupe important et d'entropie faible et un groupe peu important et d'entropie élevée, un partage équilibré des élèves de TES en deux groupes comparables d'entropie faible et d'entropie élevée et enfin une répartition des étudiants également en deux groupes, le premier d'entropie faible et d'effectif peu important, le second d'entropie élevée et rassemblant une large majorité de cette sous-population (cf. figure 1). Les élèves de CM seraient alors caractérisés par des choix plus rigides que ceux des étudiants.

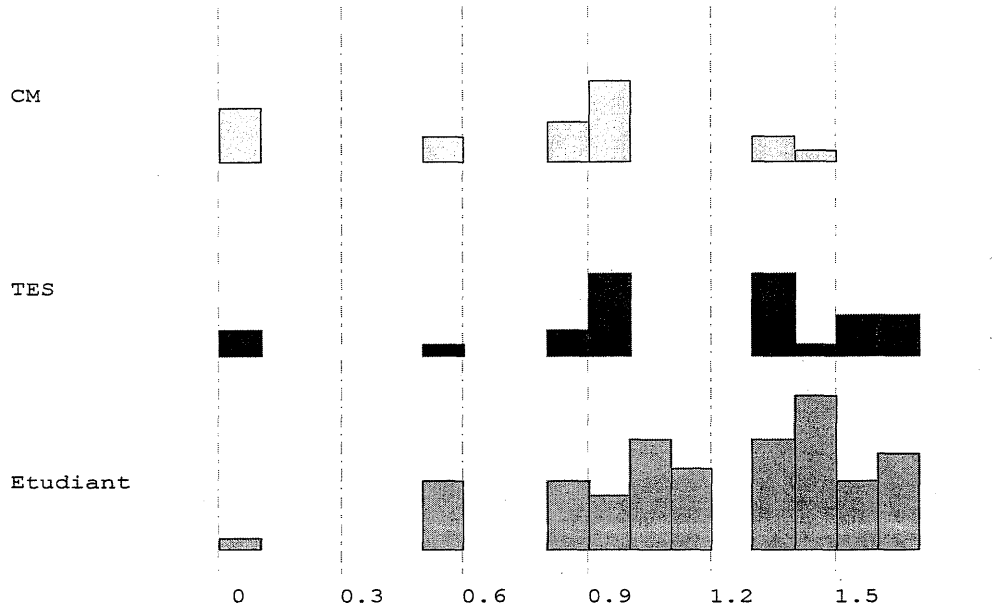

Figure 1. Distribution de l'entropie selon le statut scolaire des sujets[4]

L'analyse de la variance permet d'évaluer l'influence de l'appartenance à un sous-groupe de la population sur la distribution correspondante de l'entropie.

3.1 ANALYSE DE LA VARIANCE

L'analyse de la variance révèle une absence significative d'homogénéité des variances inter-groupes et intra-groupe au seuil 1 % (F = 8.899). Cependant, ce résultat[5] est invalidé du fait que les variances des distributions de l'entropie selon le groupe de CM et celui des étudiants ne peuvent pas être considérées comme homogènes[6] au risque 5 %. Même si l'analyse de la variance était recevable, elle ne nous aurait permis d'établir que des comparaisons (les élèves de CM sont plus rigides que les étudiants)

---

[4] Histogramme réalisé à l'aide du logiciel ADSO : A. Dubus, ADSO 3, Trois-Monts, Trigone, CEDEP, 1995.

[5] Le test de Kolmogorov donne respectivement pour la distribution de l'entropie des CM : 0.172, des TES : 0.140 et pour celle des étudiants : 0.113.

[6] F = 1.771, pour un couple (17,59).

sans nous indiquer des entropies attribuées aux catégories de sujets ni permettre une définition commune à ces sous-populations de ce que représentent une entropie élevée et une entropie faible. Il ne permet pas de définir une valeur « seuil » de l'entropie séparant statistiquement les entropies observées sur le groupe des élèves de CM des entropies observées sur le groupe des étudiants.

3.2 ANALYSE IMPLICATIVE DE DONNÉES

Nous avons retenu l'outil constitué par l'implication statistique [1] parce qu'elle ne requiert aucune hypothèse quant à la distribution de la variable numérique, mais de plus fournit une signification aux éléments descriptifs préétablis, singuliers à chacune des distributions étudiée séparément. La liste des liens implicatifs significatifs est ici vide[7]. Cette liste est, le plus souvent, peu exploitable, car elle est rarement lisible dès que le nombre de modalités observées de l'une ou l'autre des variables devient important.

Nous avons donc regroupé la variable entropie dans l'espoir de voir apparaître des liens statistiques implicatifs possédant une intensité intéressante. Alors, soit on teste les regroupements pour l'ensemble des sous-populations, soit on les teste sous-population par sous-population. Puisque le traitement algorithmique peut être transposé facilement d'un cas à l'autre, nous exposerons tout d'abord l'algorithme général, qui prend en charge le tableau initial dans sa complexité. Nous reviendrons ensuite sur l'autre possibilité de traitement.

4. DESCRIPTION DE L'ALGORITHME ET MISE EN ŒUVRE SUR L'EXEMPLE TRAITÉ

4.1 FINALITÉ DE L'ALGORITHME

Soit f et g deux variables (dans l'exemple traité, f est la variable « statut » et g la variable « entropie »).

On suppose que le nombre de modalités initiales observées de f et de g est au moins égal à 2.

On désigne par $a_1, a_2,\ldots, a_i,\ldots,$ les modalités de f. Leur nombre initial est de n_0.

On désigne par $b_1, b_2,\ldots, b_j,\ldots,$ les modalités de g. Leur nombre initial est de m_0.

La finalité de l'algorithme est de réduire au maximum le nombre des valeurs de g et de f tout en augmentant l'intensité des liens implicatifs significatifs entre valeurs de g et valeurs de f. Cet algorithme a pour but d'exhiber les regroupements de *taille maximale* des modalités observées de g et de f, pour lesquels il existe une implication

[7] Dans d'autres cas, elle pourrait être illisible parce que trop longue. Nous avons rencontré ce cas en particulier lors de l'étude de croisement de variables numériques [6].

statistique d'intensité maximale, du type : $\bigcup_{\substack{1\le i\le n_0 \\ 0\le l\le n_0-i}} \{a_{i+l}\} \Rightarrow \bigcup_{\substack{1\le j\le m_0 \\ 0\le k\le m_0-1}} \{b_{j+k}\}$, où les regroupements effectués ne se chevauchent pas.

Nous présenterons l'algorithme dans le cas où la réduction recherchée est celle des valeurs de g.

4.2 Liste des contraintes

Les contraintes qui vont guider la mise en place de cet algorithme sont les suivantes :

1. seuls les regroupements de modalités connexes de l'entropie, en tant que variable numérique sont à envisager ;

2. seuls les regroupements n'ayant aucun élément commun pourront être effectués simultanément ;

3. le gain ou la perte occasionnée par le nouveau regroupement sera déterminé en fonction de l'intensité des nouveaux liens implicatifs. En particulier, l'algorithme s'appuie sur l'opération suivante : l'intérêt du regroupement de k (pour l compris entre 1 et m_0) modalités b_l et $b_{l+1},\ldots, b_{l+k-1}$ de g dépend de la comparaison de l'intensité des liens implicatifs associés respectivement aux implications $a_i \Rightarrow b_l$ et $a_i \Rightarrow b_{l+1}$ ... et $a_i \Rightarrow b_{l+k-1}$ à l'intensité de l'implication statistique $a_i \Rightarrow (b_l$ ou b_{l+1} ... ou $b_{l+k-1})$.

Ainsi, l'intérêt du regroupement de k modalités $b_l, b_{l+1},\ldots, b_{l+k-1}$ de g est-il lié :

- pour toutes les modalités de f, soit pour toutes les valeurs de i comprises entre 1 et n_0, aux signes des *indices*[8] d'implication $q_{i,k,l}$ associés aux implications statistiques $a_i \Rightarrow (b_l$ ou b_{l+1} ... ou $b_{l+k-1})$. Un signe négatif traduit une intensité d'implication supérieure à 0.5.

- pour toutes les modalités de f, soit pour toutes les valeurs de i comprises entre 1 et n_0, à la comparaison de ces nouveaux indices d'implication $q_{i,k,l}$ aux indices d'implication $q_{i,l}, q_{i,l+1},\ldots, q_{i,l+k-1}$ respectivement associés aux implications $a_i \Rightarrow b_l$, $a_i \Rightarrow b_{l+1},\ldots, a_i \Rightarrow b_{l+k-1}$. Toute augmentation de l'intensité correspond à une diminution de l'indice d'implication.

---

[8] Nous rappelons que l'indice q associé à l'implication $a \Rightarrow b$, est égal à : $\dfrac{n_{a\wedge b'} - \frac{n_a n_{b'}}{n}}{\sqrt{\frac{n_a n_{b'}}{n}}}$ où n_a est l'effectif observé de la modalité a et où $n_{b'}$ est celui du complémentaire de la modalité b. L'intensité φ d'implication est alors $\frac{1}{\sqrt{2\pi}} \int_q^\infty e^{-\frac{t^2}{2}} dt$.

Par conséquent, l'expression $F_{i,k,l} = |q_{i,k,l}|\ (\inf\{q_{i,l}, q_{i,\,l+1},..., q_{i,l+k-1}\} - q_{i,k,l})$ est positive si le regroupement proposé traduit un gain d'intensité implicative, négative sinon.

Le gain que représente le regroupement des k valeurs de g sur l'ensemble des modalités de f est lié au signe de l'expression :

$$F_{k,l} = \sup_{1\le i\le n_0} |\, q_{i,j,l}\,|\ (\inf\{q_{i,1}, q_{i,l+1}, ..., q_{i,l+k-1}\} - q_{i,k,l})$$

4. de même, le gain occasionné par le nouveau regroupement sera déterminé en fonction de la perte d'information que représente ce regroupement, c'est-à-dire, puisqu'il s'agit toujours de modalités numériques, en fonction de l'amplitude de l'intervalle que représente ce nouveau regroupement. Pour mesurer l'intérêt d'un regroupement de k valeurs connexes de g, on prendra en compte l'amplitude de l'intervalle $[b_l, b_{l+k-1}]$ associé au regroupement $\{b_l,b_{l+1},...,b_{l+k-1}\}$, si les modalités b_j sont des valeurs discrètes. En revanche, si les b_j sont des intervalles de $\mathbb{R}$, on prend en compte la différence entre les milieux des intervalles b_l et b_{l+k-1}.

L'indice d'intérêt final associé à un regroupement $b_l, b_{l+1},..., b_{l+k-1}$ est donc :

$$I_{k,l} = \frac{F_{k,l}}{(b_{l+k-1} - b_l)}$$

Ainsi, $F_{k,l}$ étant donné, $I_{k,l}$ sera d'autant plus grand que l'amplitude sera faible.

4.3. L'ALGORITHME

Chaque étape de l'algorithme correspond à un essai de réduction d'une liste des valeurs de g, c'est-à-dire à un essai de réduction du nombre des lignes (ou des colonnes) d'un tableau croisé.

Chaque étape de l'algorithme se définit ainsi par la donnée d'un nombre m de valeurs distinctes de g : $b_1, b_2,..., b_m$ $(m \le m_0)$.

Sont associés à cette liste initiale, pour $1 \le i \le n_0$, $1 \le j \le m$:

- les effectifs $E_{i,j}$ correspondants à ces différentes valeurs, au sein de la population ;
- les indices des implications statistiques $q_{i,j}$ associés à chacune des implications statistiques $a_i \Rightarrow b_j$.

Pour toutes les valeurs successives de k, pour $2 \le k \le m$, l'intérêt (quant au gain en intensité implicative) de chacun des $m - k+1$ regroupements de k valeurs connexes de g parmi les m valeurs de g est testé.

Pour $2 \leq k \leq m$, pour $1 \leq l \leq m - k+1$, chaque regroupement des $m - k+1$ valeurs de g : $b_l, b_{l+1},..., b_{l+k-1}$ sera noté $B_{k,l}$. Pour chacun de ces regroupements on calcule alors l'indicateur numérique $I_{k,l}$.

Si $I_{k,l}$ est négatif ou nul, le regroupement $B_{k,l}$ est déclaré non intéressant et n'est pas à considérer.

1. Si, à k fixé, les regroupements $B_{k,l}$ sont tous déclarés non intéressants, la valeur de k est incrémentée, tant que k est inférieur à m. Si $k = m$ l'algorithme s'arrête.
2. Sinon, les indices $I_{k,l}$ positifs sont ordonnés. Les regroupements $B_{k,l}$ d'indices positifs sont effectués, dans l'ordre décroissant des $I_{k,l}$, à condition que ceci soit possible, c'est-à-dire :

- que le regroupement à effectuer soit compatible avec les regroupements déjà effectués ;
- que deux regroupements de même indice à effectuer puissent l'être simultanément.

Si aucun des regroupements intéressants n'est réalisable, la valeur de k est incrémentée tant que k est inférieur à m. Si $k = m$, l'algorithme s'arrête.

Sinon, les p regroupements $B_{k,l}$ de k valeurs intéressantes et réalisables sont désormais identifiés par les valeurs numériques β_l comme suit :

$$\beta_l = \frac{(b_l + b_{l+1} + ... b_{l+k-1})}{k}$$

Les p valeurs numériques β_l viennent se substituer aux $p \times k$ valeurs de g ainsi regroupées. Le nombre de valeurs distinctes de g devient $m - pk$. Une nouvelle étape de l'algorithme est amorcée, tant que le nombre de valeurs distinctes de g reste supérieure à 1. Dans le cas contraire, l'algorithme s'arrête.

4.4 EXEMPLE

Dans l'exemple présenté, les variables f et g sont les variables « niveau scolaire » et « entropie ». Les modalités de f sont au nombre de trois, celles de g sont initialement au nombre de dix : $n_0 = 3$; $m_0 = 10$.

La première étape de l'algorithme suppose donc $m = 10$. La liste initiale des 10 valeurs de g, ainsi que, pour $1 \leq i \leq 3$ et $1 \leq j \leq 10$, les valeurs des effectifs $E_{i,j}$ et des indices $q_{i,j}$ se lit sur le tableau suivant.

Tableau 3. Fonctionnement de l'algorithme. Initialisation

Entropie	Niveau scolaire		
	CM	TES	Étudiants
$.b_1 = 0.000$	$E_{1,1} = 4$ $q_{1,1} =$- 0,675	$E_{2,1} = 2$ $q_{2,1} =$ -0,075	$E_{3,1} = 1$ $q_{3,1} =$ 0,417
$.b_2 = 0.543$	$E_{1,2} = 2$ $q_{1,2} =$ -0,144	$E_{2,2} = 1$ $q_{2,2} =$ 0,187	$E_{3,2} = 5$ $q_{3,2} =$ -0,040
$.b_3 = 0.813$	$E_{1,3} = 3$ $q_{1,3} =$ -0,307	$E_{2,3} = 2$ $q_{2,3} =$ 0,076	$E_{3,3} = 5$ $q_{3,3} =$ 0,120
$.b_4 = 0.954$	$E_{1,4} = 6$ $q_{1,4} =$ -0.815	$E_{2,4} = 6$ $q_{2,4} =$ -0.497	$E_{3,4} = 4$ $q_{3,4} =$ 0.761
$.b_5 = 1.000$	$E_{1,5} = 0$ $q_{1,5} =$ 0,347	$E_{2,5} = 0$ $q_{2,5} =$ 0,400	$E_{3,5} = 8$ $q_{3,5} =$ -0,443
$.b_6 = 1.061$	$E_{1,6} = 0$ $q_{1,6} =$ 0,257	$E_{2,6} = 0$ $q_{2,6} =$ 0,297	$E_{3,6} = 6$ $q_{3,6} =$ -0,329
$.b_7 = 1.298$	$E_{1,7} = 2$ $q_{1,7} =$ 0,211	$E_{2,7} = 6$ $q_{2,7} =$ -0,497	$E_{3,7} = 8$ $q_{3,7} =$ 0,198
$.b_8 = 1.405$	$E_{1,8} = 1$ $q_{1,8} =$ 0,327	$E_{2,8} = 1$ $q_{2,8} =$ 0,450	$E_{3,8} = 11$ $q_{3,8} =$ -0,463
$.b_9 = 1.500$	$E_{1,9} = 0$ $q_{1,9} =$ 0,347	$E_{2,9} = 3$ $q_{2,9} =$ -0,238	$E_{3,9} = 5$ $q_{3,9} =$ -0,040
$.b_{10} =1.561$	$E_{1,10} = 0$ $q_{1,10} =$ 0,438	$E_{2,10} = 3$ $q_{2,10} =$ -0,139	$E_{3,10} = 7$ $q_{3,10} =$ -0,152

Pour la première étape, k vaudra tout d'abord 2. Le tableau 4 présente les 9 regroupements, notés $B_{2,l}$ pour $1 \leq l \leq 9$, des 10 valeurs initiales de l'entropie, deux à deux, ainsi que les indicateurs $F_{i,j,l}$, $F_{k,l}$ et $I_{2,l}$.

Tableau 4. Fonctionnement de l'algorithme : test de l'intérêt du regroupement des premières valeurs de l'entropie

Entropie	Niveau scolaire			$F_{k,l}$
	CM	TES	Étudiants	$I_{k,l}$
$B_{2,1} = \{b_1\} \cup \{b_2\}$	$q_{1.2.1} = -0.856$ $F_{1.2.1} = 0.154$	$q_{2,2,1} = 0.117$ $F_{2.2.1} = -0.022$	$q_{3,2,1} = 0.394$ $F_{3.2.1} = -0.171$	$F_{2.1} = 0.154$ $I_{2,1} = 0.284$
$B_{2,2} = \{b_2\} \cup \{b_3\}$	$q_{1.2.2} = -0.474$ $F_{1.2.2} = 0.079$	$q_{2,2,2} = 0.276$ $F_{2.2.2} = -0.056$	$q_{3,2,2} = 0.084$ $F_{3.2.2} = -0.010$	$F_{2.2} = 0.079$ $I_{2,2} = 0.293$
$B_{2,3} = \{b_3\} \cup \{b_4\}$	$q_{1.2.3} = -1.205$ $F_{1.2.3} = 0.469$	$q_{2.2.3} = -0.445$ $F_{2.2.3} = -0.023$	$q_{3,2,3} = 0.941$ $F_{3.2.3} = -0.773$	$F_{2.3} = 0.469$ $I_{2,3} = 3.326$
$B_{2,4} = \{b_4\} \cup \{b_5\}$	$q_{1.2.4} = -0.476$ $F_{1.2.4} = -0.162$	$q_{2.2.4} = -0.082$ $F_{2.2.4} = -0.034$	$q_{3,2,4} = 0.313$ $F_{3.2.4} = -0.236$	$F_{2.4} = -0.034$ $I_{2.4} = -0.742$
$B_{2,5} = \{b_5\} \cup \{b_6\}$	$q_{1,2,5} = 0.627$ $F_{1.2.5} = -0.232$	$q_{2,2,5} = 0.724$ $F_{2.2.5} = -0.309$	$q_{3.2.5} = -0.801$ $F_{3.2.5} = 0.287$	$F_{2.5} = 0.287$ $I_{2,5} = 4.705$
$B_{2,6} = \{b_6\} \cup \{b_7\}$	$q_{1,2,6} = 0.501$ $F_{1.2.6} = -0.145$	$q_{2.2.6} = -0.190$ $F_{2.2.6} = -0.058$	$q_{3.2.6} = -0.154$ $F_{3.2.6} = -0.027$	$F_{2.6} = -0.027$ $I_{2.6} = -0.114$
$B_{2,7} = \{b_7\} \cup \{b_8\}$	$q_{1,2,7} = 0.590$ $F_{1.2.7} = -0.223$	$q_{2.2.7} = -0.043$ $F_{2.2.7} = -0.019$	$q_{3.2.7} = -0.296$ $F_{3.2.7} = -0.050$	$F_{2.7} = -0.019$ $I_{2.7} = -0.181$
$B_{2,8} = \{b_8\} \cup \{b_9\}$	$q_{1.2.8} = 0.716$ $F_{1.2.8} = -0.279$	$q_{2,2,8} = 0.216$ $F_{2.2.8} = -0.098$	$q_{3.2.8} = -0.528$ $F_{3.2.8} = 0.034$	$F_{2.8} = 0.034$ $I_{2,8} = 0.361$
$B_{2.9} = \{b_9\} \cup \{b_{10}\}$	$q_{1,2,9} = 0.825$ $F_{1,2,9} = -0.395$	$q_{2.2.9} = -0.397$ $F_{2,2,9} = 0.063$	$q_{3.2.9} = -0.201$ $F_{3,2,9} = 0.010$	$F_{2.9} = 0.063$ $I_{2,9} = 1.037$

Six valeurs $I_{k,l}$ seulement sont positives : $I_{2,1}$, $I_{2,2}$, $I_{2,3}$, $I_{2,5}$, $I_{2,8}$, $I_{2,9}$.

Ces valeurs ordonnées par ordre décroissant sont : $I_{2,5}$ (4,705), $I_{2,3}$ (3,326), $I_{2,9}$ (1,037), $I_{2,8}$(0,361), $I_{2,2}$(0,293) et $I_{2,1}$ (0,284). 4 regroupements sont possibles à effectuer : $B_{2,5}$, $B_{2,3}$, $B_{2,9}$ et $B_{2,1}$.

Les nouvelles valeurs prises par g sont au nombre de 6 : $b_1 = (0.000+0.543)/2$, $b_2 = (0.543+0.813)/2$, $b_3 = (1.000+1.061)/2$, $b_4 = 1.298$, $b_5 = 1.405$, $b_6 = (1.500+1.561)/2$.

On continue, en reprenant $k = 2$.

Soit :

Tableau 5. Deuxième pas

Entropie	Niveau scolaire		
	CM	TES	Étudiants
$.b_1 = 0.2715$	$E_{1,1} = 6$ $q_{1,1}$ =-0.856	$E_{2,1} = 3$ $q_{2,1}$= 0.117	$E_{3,1} = 6$ $y_{3,1}$ = 0,395
$.b_2 = 0.678$	$E_{1,2} = 9$ $q_{1,2}$ = -1,205	$E_{2,2} = 8$ $q_{2,2}$ = -0,445	$E_{3,2} = 9$ $y_{3,2}$= 0,941
$.b_3 = 1.0305$	$E_{1,3} = 0$ $q_{1,3}$ = 0,627	$E_{2,3} = 0$ $q_{2,3}$ = 0,724	$E_{3,3} = 14$ $y_{3,3}$ = -0,801
$.b_4 = 1.298$	$E_{1,4} = 2$ $q_{1,4}$ = 0,211	$E_{2,4} = 6$ $q_{2,4}$ = -0.497	$E_{3,4} = 8$ $y_{3,4}$ = 0,198
$.b_5 = 1.405$	$E_{1,5} = 1$ $q_{1,5}$ = 0,327	$E_{2,5} = 1$ $y_{2,5}$ = 0,450	$E_{3,5} = 11$ $y_{3,5}$ = -0,463
$.b_6 = 1.5305$	$E_{1,6} = 0$ $q_{1,6}$ = 0,825	$E_{2,6} = 6$ $y_{2,6}$ = -0,397	$E_{3,6} = 12$ $y_{3,6}$ = -0,201

Les regroupements deux à deux sont de nouveau testés :

Entropie	Niveau scolaire			$F_{k,l}$ $I_{k,l}$
	CM	TES	Étudiants	
$B_{2,1} = \{b_1\} \cup \{b_2\}$	$q_{1,2,1}$=-2,367 $F_{1,2,1}$=2,750	$q_{2,2,1}$=-0,357 $F_{2,2,1}$=-0.031	$q_{3,2,1}$=1,522 $F_{3,2,1}$=-1,716	$F_{2,1}$=2,750 $I_{2,1}$=3,382
$B_{2,2} = \{b_2\} \cup \{b_3\}$	$q_{1,2,2}$=-0.488 $F_{1,2,2}$=-0,350	$q_{2,2,2}$=0.501 $F_{2,2,2}$=-0,474	$q_{3,2,2}$=-0.050 $F_{3,2,2}$=-0,037	$F_{2,2}$=-0,037 $I_{2,2}$=-0,053
$B_{2,3} = \{b_3\} \cup \{b_4\}$	$q_{1,2,3}$=0,924 $F_{1,2,3}$=-0,659	$q_{2,2,3}$=0,257 $F_{2,2,3}$=-0,194	$q_{3,2,3}$=-0,669 $F_{3,2,3}$=-0,089	$F_{2,3}$=-0,089 $I_{2,3}$=-0,165
$B_{2,4} = \{b_4\} \cup \{b_5\}$	$q_{1,2,4}$=0,590 $F_{1,2,4}$=-0,223	$q_{2,2,4}$=-0,043 $F_{2,2,4}$=-0,019	$q_{3,2,4}$=-0,296 $F_{3,2,4}$=-0,050	$F_{2,4}$=-0,019 $I_{2,4}$=-0,090
$B_{2,5} = \{b_5\} \cup \{y_6\}$	$q_{1,2,5}$=1,263 $F_{1,2,5}$=-1,183	$q_{2,2,5}$=0.072 $F_{2,2,5}$=-0,034	$q_{3,2,5}$=-0,737 $F_{3,2,5}$=0.202	$F_{2,5}$=0,202 $I_{2,5}$=0,804

Ainsi, deux regroupements vont être effectués à l'issue de cette étape : $B_{2,1}$ et $B_{2,5}$.

On obtient enfin (voir tableaux 6 et 6 bis), une partition des valeurs prises par l'entropie à deux classes. Sur chacun de ces intervalles, l'intensité q de l'implication statistique entre la catégorie « CM » d'une part et « étudiant » d'autre part satisfait aux exigences émises (seuil de risque inférieur à 10 %).

Tableau 6. Regroupements finaux des valeurs de l'entropie.
Indices d'implications statistiques entre catégories de sujets et intervalles d'entropie

Entropie	0,000 à 0,954	1,000 à 1,561
CM	$E_{1,1}$= 15 $q_{1,1}$ =-2,367	$E_{1,2}$=3 $q_{1,2}$= 2,886
TES	$E_{2,1}$=11 $q_{2,1}$= -0,357	$E_{2,2}$=13 $q_{2,2}$= 0,436
Étudiant	$E_{3,1}$=15 $q_{3,1}$=1,522	$E_{3,2}$=45 q=-1,856

Tableau 6 bis. Intensités des implications statistiques entre catégories de sujets et regroupements des valeurs de l'entropie

Entropie	0,000 à 0,954	1,000 à 1,561
CM	99%	ns
TES	ns	ns
Étudiant	ns	96%

L'algorithme permet ainsi d'opposer les élèves de CM aux étudiants de licence. Les élèves de CM sont caractérisés par une entropie faible correspondant à des choix rigides. Les étudiants sont caractérisés par des choix plus diversifiés. Ces opérations permettent de donner une définition d'une entropie faible : c'est une entropie inférieure à 1, et une entropie forte est une entropie supérieure ou égale à 1. Les élèves de CM sont caractérisés par des choix recouvrant au maximum deux des trois conceptions du hasard reconstruites, tandis que les étudiants sont caractérisés par des choix englobant les trois mêmes conceptions (voir Tableau 1). L'algorithme a été ensuite mené sur les valeurs de f, mais aucun regroupement ne s'avère intéressant.

CONCLUSION : AUTRE FONCTIONNEMENT POSSIBLE DE L'ALGORITHME

Il est aussi possible de rechercher, comme nous l'avons signalé au début de l'exposé opérationnel, à faire fonctionner cet algorithme, catégorie de sujets par catégorie de sujets, c'est-à-dire par valeur fixée de f. Dans ce cas, on recherche par exemple, quel regroupement de l'entropie conduit à des implications statistiques du type : « CM » $\Rightarrow$ « une entropie comprise entre a_1 et a_n », avec une intensité intéressante, puis quels sont ceux qui permettent d'écrire des implications statistiques du type : « TES » $\Rightarrow$ « entropie comprise entre a_i et a_m », etc.

Le traitement permet d'accéder ainsi à d'autres renseignements. Mais rien n'assure que les regroupements intéressants retenus en dernière instance se recoupent. La mise en œuvre de l'algorithme sert moins dans ce cas à révéler des oppositions entre groupes de sujets qu'à transcrire des lignes de force implicatives.

BIBLIOGRAPHIE

[1] GRAS, R., *L'implication statistique, nouvelle méthode exploratoire de données*, Grenoble, La Pensée Sauvage éditions, 1996.

[2] GRAS, R., LAHRER, A., « L'implication statistique, une nouvelle méthode d'analyse des données », *Mathématiques, Informatique et Sciences humaines*, 120, 1993, p. 5-31.

[3] HOEL, P., *Statistique mathématique*, tome II, Paris, Armand Colin, 1991.

[4] LAHANIER-REUTER, D., *Étude de conceptions du hasard : approche épistémologique, didactique et expérimentale en milieu universitaire*, Thèse de doctorat, Université de Rennes I, 1998.

[5] LAHANIER-REUTER, D., *Conceptions du hasard et enseignement des probabilités et des statistiques*, Paris, Presses Universitaires de France, 1999.

[6] LAHANIER-REUTER, D., « Exemple d'une nouvelle méthode d'analyse de données : l'analyse implicative », *Carrefours de l'éducation*, 2000.

[7] LERMAN, I.C., GRAS, R., ROSTAM, H., « Élaboration et évaluation d'un indice d'implication pour les données binaires », *Mathématiques, Informatique et Sciences humaines*, 74, 1981, p. 5-35.